# Recognizing Food Places in Egocentric Photo-Streams Using Multi-Scale Atrous Convolutional Networks and Self-Attention Mechanism

Sarker, Md Mostafa Kamal; Rashwan, Hatem A.; Akram, Farhan; Talavera, Estefania; Banu, Syeda Furruka; Radeva, Petia; Puig, Domenec

*Published in:*
IEEE Access

*DOI:*
10.1109/ACCESS.2019.2902225



*Publication date:*
2019

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 23, 2019, accepted February 18, 2019, date of publication March 20, 2019, date of current version April 5, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2902225*

# Recognizing Food Places in Egocentric Photo-Streams Using Multi-Scale Atrous Convolutional Networks and Self-Attention Mechanism

**MD. MOSTAFA KAMAL SARKER[1], HATEM A. RASHWAN[1], FARHAN AKRAM[2], ESTEFANIA TALAVERA[3], SYEDA FURRUKA BANU[4], PETIA RADEVA[5], AND DOMENEC PUIG[1]**

[1]Department of Computer Engineering and Mathematics, Universitat Rovira i Virgili, 43007 Tarragona, Spain
[2]Imaging Informatics Division, Bioinformatics Institute, Singapore 138671
[3]Bernoulli Institute, University of Groningen, 729700 Groningen, The Netherlands
[4]ETSEQ, Universitat Rovira i Virgili, 43007 Tarragona, Spain
[5]Department of Mathematics and Computer Science, Universitat de Barcelona, 08007 Barcelona, Spain

Corresponding author: Md. Mostafa Kamal Sarker (mdmostafakamal.sarker@urv.cat)

This work was supported in part by the program Marti Franques under the agreement between Universitat Rovira Virgili and Fundacio Catalunya La Pedrera under Project TIN2015-66951-C2, Project SGR 1742, and Project CERCA, in part by the Nestore Horizon2020 SC1-PM-15-2017 under Grant 769643, in part by the EIT Validithi, in part by the ICREA Academia 2014, and in part by the NVIDIA Corporation.

**ABSTRACT** Wearable sensors (e.g., lifelogging cameras) represent very useful tools to monitor people's daily habits and lifestyle. Wearable cameras are able to continuously capture different moments of the day of their wearers, their environment, and interactions with objects, people, and places reflecting their personal lifestyle. The food places where people eat, drink, and buy food, such as restaurants, bars, and supermarkets, can directly affect their daily dietary intake and behavior. Consequently, developing an automated monitoring system based on analyzing a person's food habits from daily recorded egocentric photo-streams of the food places can provide valuable means for people to improve their eating habits. This can be done by generating a detailed report of the time spent in specific food places by classifying the captured food place images to different groups. In this paper, we propose a self-attention mechanism with multi-scale atrous convolutional networks to generate discriminative features from image streams to recognize a predetermined set of food place categories. We apply our model on an egocentric food place dataset called ''EgoFoodPlaces'' that comprises of 43 392 images captured by 16 individuals using a lifelogging camera. The proposed model achieved an overall classification accuracy of 80% on the ''EgoFoodPlaces'' dataset, respectively, outperforming the baseline methods, such as VGG16, ResNet50, and InceptionV3.

**INDEX TERMS** Food places recognition, scene classification, self-attention model, atrous convolutional networks, egocentric photo-streams, visual lifelogging.

## I. INTRODUCTION

Overweight and obesity yield many major risk factors for chronic diseases, including diabetes, cardiovascular diseases and cancer. According to the statistics given by WHO,[1] the obesity rate has nearly tripled since 1975. In 2016, more than 1.9 billion adults with age 18 years and older were counted overweight through the world, out of which 650 million were obese [1], [2]. Comparing the death reason of people shows that overweight and obesity kill more people than underweight and malnutrition [3]. Therefore, the concern of the preventing obesity is highly demanding in developed countries. On the other hand, the cost of health services caused by overweight and obesity are increasing for the government every year to billions of dollars [4]. For example, the obesity medical cost in Europe was estimated at around

The associate editor coordinating the review of this manuscript and approving it for publication was Ah Hwee Tan.

[1]http://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

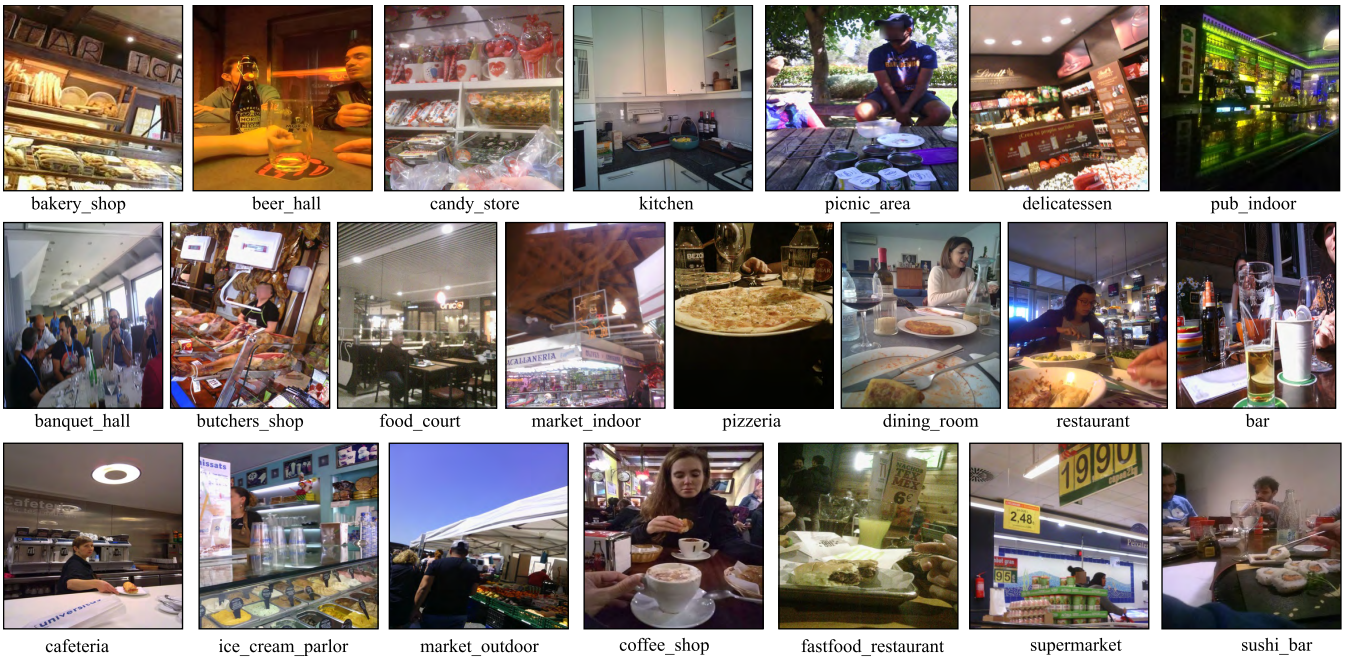



**FIGURE 1.** Examples of food places collected from the EgoFoodPlaces image dataset.

€81 billion per year in 2012. In keeping with the WHO estimates on obesity expenditure, this was 2%-8% of the total national expenditure in the 53 European countries [5].

Food environment, adverse reactions to food, nutrition, and physical activity patterns are relevant aspects for the health care professional to consider when treating obesity. Recent studies have shown that 12 cancers are directly linked to overweight and obesity[2]. The food that we eat, how active we are and how much we weigh have a direct influence on our health. Thus, by observing unhealthy diet patterns, we can create a healthy diet plan that can play a major role in our fight against obesity and being overweight. Therefore, diet patterns are important key factors that have to be analyzed for preventing overweight and obesity.

Conventional nutrition diaries are not good enough for tracking the lifestyle and food patterns properly, since they need a huge amount of human interaction. Nowadays, mobiles phones are also used to keep track of ones diet by keeping a record of food intake and their respective calories. However, this is done by taking the photos of the dishes, which can make people uncomfortable[3]. For this reason, we need an automatic system that can correctly record the user food patterns and help to analyze the lifestyle and nutrition as well. To track the food patterns, we need to answer about three questions: where, how long and with whom the person is eating. These answers can discover the details of people nutritional habits, which can help to improve their healthy lifestyle and prevent the overweight and obesity. In this work by analyzing daily user information captured by a wearable camera, we focus on the places or environment that users are commonly eating in, which is also called ''food places''.

Recording daily user information by the traditional camera is difficult. Therefore, we prefer to use wearable cameras, such as life-logging camera, being able to collect daily user information (see Figure 1). These cameras are capable of frequently and continuously capturing images that record visual information of our daily life known as ''visual life-logging''. It can collect a huge number of images by non-stop image collection capacity (1-4 per minute, 1k-3k (1k = 1000) per day and 500k-1000k per year). These images can create a visual diary with activities of the person life with unprecedented details [6]. The analysis of egocentric photo-streams (images) can improve the people lifestyle by analyzing social pattern characterization [7] and social interactions [8], as well as generating storytelling of first-person days [6]. In addition, the analysis of these images can greatly affect human behaviors, habits, and even health [9]. One of the personal tendencies of people is food events that can badly affect their health. For instance, some people get hungrier if they continuously see and smell food, consequently they end up eating more [10], [11]. Also, it is well-known that people going to shop hungry, buy more and less healthy food. Thus, monitoring the duration of food intake and the time people spend in food-related environment can help them get aware of their habits and improve their nutritional behavior.

The motivation behind this research is two-fold. Firstly, using a wearable camera is to capture images related to food places, where the users are engaged within foods (see Figure 1). Consequently, these images of visual life-logging

[2]https://www.wcrf.org/int/blog/articles/2018/05/blueprint-beat-cancer
[3]https://www.redbookmag.com/body/healthy-eating/advice/g614/lose-weight-apps-tools/

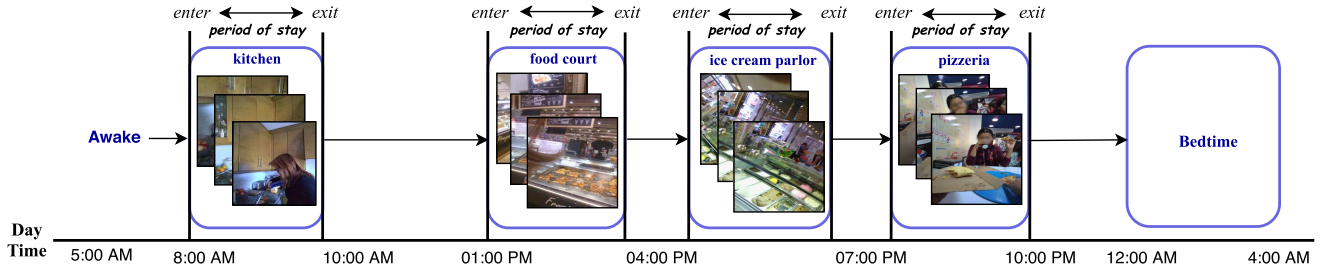



**FIGURE 2.** Examples of daily log that shows time spent in different food places.

can give a unique opportunity to work on food pattern analysis from the individual's viewpoint. Secondly, the analysis of everyday information (entering, exiting and time of stay as shown in Figure 2) of visited food places can enable a novel healthcare approach that can help to manage better diseases related to nutrition, like obesity, diabetes, heart diseases, and cancer.

This work is a progression of our previous work proposed in [12] and our main contributions can be summarized as follows:

- Design and development of a novel attention-based deep network based on the multi-scale Atrous convolutional networks [12], called MACNet with self-attention (MACNet+SA) for improving classification rate of food places.
- Application of the MACNet+SA model to treat a sequence of images for food events analysis.

The paper is organized as follows. Section 2 discusses the related works of places or scene classification. The proposed attention-based deep network architecture is described in Section 3. The experimental results and discussions are illustrated in Section 4. Finally, section 5 shows the conclusions and future work.

## II. RELATED WORKS

Early work of places or scene recognition in conventional images has been discussed in the literature by applying classical approaches [13]–[16]. The traditional scene classification methods can be classified into two main categories: generative models and discriminative models. Generative models are generally hierarchical Bayesian systems to characterize a scene, which can represent different relations in a complex scene [17]–[19]. Discriminative models are to extract dense features of an image and encode the features into a fixed length description to build a reasonable classifier for scene recognition [20], [21]. The discriminative classifiers, such as logistic regression, boosting and Support Vector Machine (SVM) were widely adopted for scene classification [22]. In [23], the authors recognized 15 different categories of outdoor and indoor scenes by computing histograms of local features of image parts. In turn, [24] proposed a scene classification method for indoor scenes (i.e., total 67 categories of scenes; 10 of them are related to food places). The method is based on a combination of local and global features of the input images.

Recently, the Convolutional Neural Networks (CNNs) have shown fruitful applications to digits recognition. CNNs have become a more powerful tool after introducing AlexNet [25] based on the large-scale dataset called "ImageNet" [26]. Afterwards, the history of CNN evolution began with many breakthroughs, such as VGG16 [27], Inception [28] and ResNet50 [29]. The era of places classification turned into new dimensions after introducing two large-scale places datasets, Places2 [30] and SUN397 [31] with millions of labeled images. The combination of using deep learning models with large-scale dataset outperforms the traditional scene classification methods [32].

An overall of the state-of-the-art places or scene classification based on deep networks has been discussed in a review article presented in [32]. However, the performance of *scene recognition* challenges shown in [32] has not achieved the same level of success as *object recognition* challenges [26]. This outcome showed the difficulty of the general classification problem between scene and object level, as a result of large different places surroundings people (e.g., 400 places in Places2 dataset [32]). Zheng *et al.* [33] proposed a probabilistic deep embedding framework for analyzing scenes by combining local and global features extracted by a CNN network. In addition, two separate networks called "Object-Scene CNNs" proposed in [34], in which a composed model of 'object net' and 'scene net' for aggregating information from the outlook of objects performs scene recognition. The two networks were pre-trained on the ImageNet dataset [26] and Places2 dataset [32], respectively. Indeed, many of deep architectures were evaluated on these datasets based on the conventional images. None of them is tested on the egocentric images that themselves represent a challenge for image analysis.

Recently, egocentric image analysis is a very promising field within computer vision for developing algorithms for understanding the first person personalized scenes. Many classifiers were used to classify 10 different categories of scenes based on egocentric videos [35]. They trained the classifiers by using One-vs-All cross-validation. In addition, a multi-class classifier with a negative-rejection technique was proposed in [36]. Both works [35], [36] considered

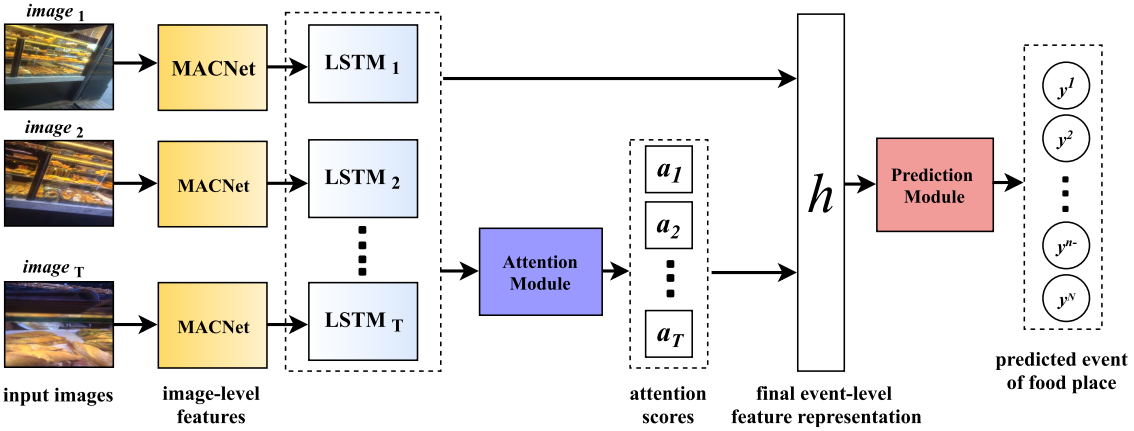



**FIGURE 3.** Architecture of our proposed attention-based model for food places classification.

only 10 categories of scenes, 2 of them are related to food places (i.e., *kitchen* and *coffee machine*). Moreover, some places related to food and type of food are classified in [37] and [38] by using conventional images from the Places2 and CuisineNet dataset [32], [38].

In our previous work [12], we introduced a deep network named "MACNet" based on atrous convolutional networks [39] for food places classification. The MACNet model is based on a pre-trained ResNet and works on images without using any time dependence [12]. In addition, food places recognition is still a challenge due to the big variety of food places environments in real-world, and the wide range of possibilities of how a scene can be captured from the person's point of view. Therefore, we re-define our problem based on the relevant temporal intervals (period of stay time). This period is divided into a set of events that is a sequence of correlated egocentric photos. A self-attention deep model will then be used to classify these events. To the best of our knowledge, this is the first work on the food places pattern classification based on an event of a stream of egocentric images in order to create intelligent tools for food-related environment monitoring.

## III. PROPOSED APPROACH

Recently, the Recurrent Neural Netwok (RNN) and attention-based models are widely used in the fields of Natural Language Processing (NLP), such as [40] for image captioning [41], for video captioning [42], and for sentiment analysis [43], [44]. In these approaches, a query vector is commonly used, which contains relevant information (i.e., in our case it is image-level features) for generating the next token in order to pick relevant parts of the input as supplementary context features. The attention models can be classified into two categories [41], namely local (hard) and global (soft) attention. The hard attention selects only a part of input, which is non-differentiable that needs a more complex algorithm, such as variance reduction or reinforcement learning to train. In turn, the soft attention is based on a softmax function to create a global decision on all parts of the input sequence. In addition, back-propagation is commonly used for training the attention models with both mechanisms in various tasks.

One of the effective soft-attention models is a self-attention mechanism [45] with no extra queries. The self-attention mechanism can easily estimate the attention scores based on a self-representation. In this work, our attention model follows the self-attention scheme, where features extraction from the input images is done using the pre-trained MACNet model. LSTM cells are used to compute the attention scores. That is done by feeding these image-level features to an attention module to generate event-level features that the prediction module uses to classify the input event.

### A. NETWORK ARCHITECTURE

The main framework of our proposed attention-based model for food places classification is illustrated in Figure 3. The proposed model consists of three major modules: features extraction, attention and prediction modules.

The feature extraction module is based on the MACNet [12] model that is fed by one input image from a food place event, see Figure 4. In MACNet [12], the input image is scaled into five different resolutions (i.e. the original image with four different resolutions with a scale value of 0.5). The original input image resolution, $I$ is $224 \times 224$ (i.e. standard input size of residual network [29]). The five scaled images are fed to five blocks of an atrous convolutional networks [39] with three different rates (in this work, we used rates $= 1$, 2, and 3) to extract the key features of the input image in a multi-scale framework. In addition, four layers (blocks) of pre-trained ResNet101 are used sequentially to extract 256, 512, 1024 and 2048 feature maps, respectively as shown in Figure 4. Each feature maps extracted by an atrous convolutional block is concatenated with the corresponding ResNet block to feed the subsequent block. Finally, the features

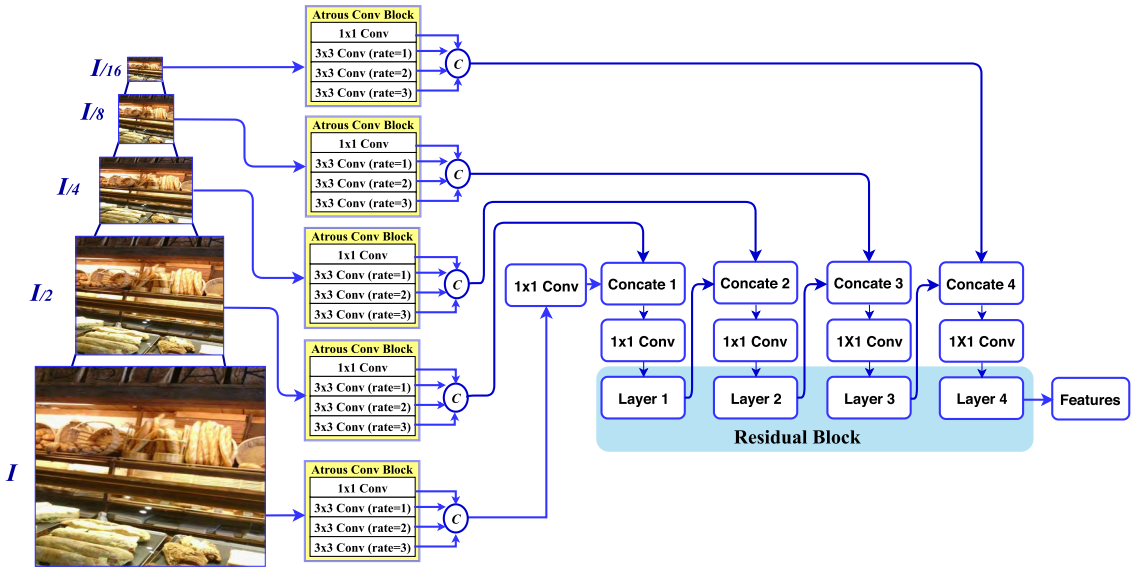



**FIGURE 4.** Architecture of our previous work, MACNet [12], for the image-level feature extraction.

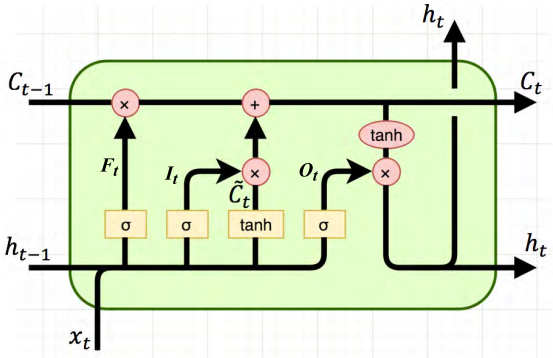



**FIGURE 5.** Standard architecture of an LSTM cell.

obtained from the fourth ResNet layer are the final features used to describe the input image.

In the second step, a Long Short-Term Memory (LSTM) unit (LSTM cell) [46] is applied designed to learn long-term dependencies features of all images per event. This unit consists of a number of LSTM cells. Figures 5 illustrates the LSTM cells properties. A classical LSTM cell consists of three sigmoid layers: a forget gate layer, an input gate layer, and an output gate layer. The three layers determine the information to flow-in and flow-out at the current time step. The mathematical definitions of these layers can be defined as:

$$F_t = \sigma(W_F.[h_{t-1}, x_t] + b_F), \quad (1)$$

$$I_t = \sigma(W_I.[h_{t-1}, x_t] + b_I), \quad (2)$$

$$O_t = \sigma(W_O.[h_{t-1}, x_t] + b_O), \quad (3)$$

where, $\sigma$ represents the sigmoid function, $x_t$ is the input features vector at time $t$, $h_{t-1}$ is the output state of the LSTM cell at the previous step at time $t-1$, $F_t$, $I_t$, and $O_t$ are the outputs of the three gates layers at time $t$, $W_j$, and $b_j$ are a weight matrix and a bias scalar for a layer, where $j$ is for $F$, $I$ or $O$ layers. For updating the cell state, the LSTM cell also needs a *tanh* layer to create a vector of new candidate values, $\tilde{C}_t$, which can be computed after the information coming from the input gate layer by:

$$\tilde{C}_t = tanh(W_C.[h_{t-1}, x_t] + b_C], \quad (4)$$

where $W_C$ and $b_C$ are a weight matrix and a bias scalar for the *tanh* layer. The old cell state, $C_{t-1}$, to the new cell state, $C_t$ can be updated by combining the outputs of the forget and the input gate layers by:

$$C_t = F_t * C_{t-1} + I_t * \tilde{C}_t \quad (5)$$

Finally, the output state of the LSTM cell is:

$$h_t = O_t * tanh(C_t). \quad (6)$$

In our model, the outputs of the MACNet model are the features extracted from the input images of an event $x_0, x_1, \cdots, x_T$. These features are fed to a set of LSTM cells, for capturing additional context dependencies features. Assume we have $T$ number of LSTM cells, $\{LSTM_1, \cdots\cdots\cdots, LSTM_T\}$, $LSTM_t \in \mathbb{R}^H$, where $T$ is the number of images and $H$ is the dimension of the extracted features vector. The output features of the LSTM cells are sequentially fed to an attention module in order to ensure that the network is able to increase its sensitivity to the important features, and suppress less useful features. The attention module will be learned how to average image-level features in a weighted manner. The weighted average is obtained by weighting each image-level features by a factor of its product with a global attention vector. The features vector of each image and the global attention vector will be trained and learned simultaneously using a standard back-propagation algorithm. In our proposed model, we use the dot product between global attention vector $V$ and image-level feature

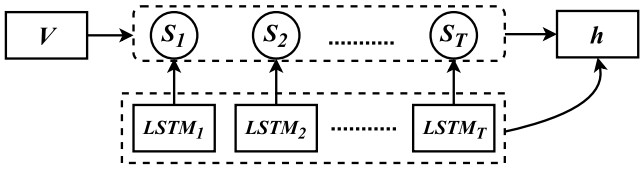



**FIGURE 6.** Global self-attention mechanism for final event-level feature representation.

$LSTM_t$ as a score of the $t$-th image. Thus this score can be computed as:

$$S_t =< V, LSTM_t > . \tag{7}$$

The global attention vector, $V \in \mathbb{R}^H$ is initialized randomly and learned simultaneously by the network. To construct image-level features for different food-places events, the global attention vector, $V$ can learn the general pattern of the event relevance of images. The architecture of the global self-attention mechanism is shown in Figure 6. Multiple information of successive images is aggregated into a single event-level vector representation with attention. The attention mechanism computes a weighted average over the combined image-level features vectors, and its main job is to compute a scalar weight to each of them. For constructing the final event-level representation, it is also not differentiated whether the images belong to the target event or any other events.

The attention module measures a score, $S_t$ for each image-level features $LSTM_t$ and normalizes it by a softmax function as follows:

$$\alpha_t = \frac{exp(S_t)}{\sum_{t=1}^{T} exp(S_t)}, \tag{8}$$

where $\alpha$ is the probabilistic heat-map. Thus, the image-level features $LSTM_t \in \mathbb{R}^H$ are then biased by the corresponding attention scores. The final event-level features, $h$ are the element-wise weighted average of all the image-level features defined as:

$$h = \sum_{t=1}^{T} \alpha_t LSTM_t, \tag{9}$$

where $h$ is the event-level features that will be used to automatically train the prediction module to predict the events of a period of stay in a food-place. There are various type of the prediction modules available in the literature. In this work, a fully connected neural network is used as a multi-label event prediction module:

$$\hat{y}^n = p(y^n|h) = \frac{1}{1 + e^{-}(w^n h + b^n)} \in [0, 1], \tag{10}$$

where $\hat{y}^n$ is the predicted label, $y^n$ is the ground-truth of the $n$-th event, $n = 1$ to $N$, $N$ is the total number of events samples, and $w^n$ and $b^n$ are the classification weight and biasing parameters, respectively, for predicting the $n$-th event. The whole model trained end-to-end by minimizing the multi-label classification loss is given by:

$$\ell = -\frac{1}{N} \sum_{n=1}^{N} E(y^n, \hat{y^n}), \tag{11}$$

where $E$ is the cross-entropy function.

## IV. EXPERIMENTAL RESULTS

### A. EGOFOODPLACES DATASET

Initially, we employed the egocentric dataset ''EgoFoodPlaces'' in our previous work [12]. However, in this work, ''EgoFoodPlaces'' was modified by adding more images. As well as, each class contains a set of events (i.e. a sequence of images) instead of still images. Our egocentric dataset, ''EgoFoodPlaces'', was constructed by 16 users using a lifelogging camera (i.e., narrative clip 2,[4] which has an image resolution of 720p and 1080p by a 8-megapixel camera with an 86-degree field of view and capable of record about 4, 000 photos or 80 minutes of 1080p video at 30fps. Figure 1 shows some example images from the ''EgoFoodPlaces'' dataset. The user fixed the camera to his/her chest from morning to night before sleeping for capturing the visual information about his/her daily environment. Thus, sets of egocentric photo-streams (events) exploring the users daily food patterns (e.g. a person spends a specific time in a food-place, such as restaurant, cafeteria, coffee shop, etc.) were captured, see Figure 2. Every frame of a photo-stream is recording first-person personalized scenes that will be used for analyzing different patterns of the user lifestyle.

However, in ''EgoFoodPlaces'', the captured images have different challenges, such as the blurriness (the effect of the user' motion), black, ambiguous and occluded images (occluded by the user hand or other body parts) during the streaming, which is not good for the entire system. All these challenges reduce the accuracy rate of a recognition system. Therefore, some pre-processing techniques are necessary to be applied to refine the collected images.

*For removing the blurry images*, we compute the blurriness amount in each image using the variance of the Laplacian. The blurriness amount is calculated by a pre-defined threshold (i.e. in this work, the threshold value is set to 500). If the variance is lower than the threshold, then the image is considered blurry. Particularly, if the image contains high variance, the image has a widespread response of both edge-like and non-edge indicating to an in-focus image. In turn, if the variance is low, the image has a tiny spread of responses specifying that the number of edges appearances in the image is very small and the image is blurred.

In turn, *for removing the black, ambiguous and occluded images* from our dataset, the K-Means clustering algorithm was used with $K = 3$ (i.e., red, green and blue). If 90% of the pixels of an image are clustered to a dominant color, we consider that the image is not informative enough, and it is eliminated from the dataset.

[4]http://getnarrative.com/

**TABLE 1.** The distribution of images per class in the EgoFoodPlaces dataset.

| | Train | | Val | | Test | | Total | |
|---|---|---|---|---|---|---|---|---|
| Classes | images | events | images | events | images | events | images | events |
| bakery shop | 356 | 36 | 108 | 11 | 128 | 13 | 592 | 60 |
| banquet hall | 420 | 42 | 150 | 15 | 146 | 15 | 716 | 72 |
| bar | 1320 | 132 | 410 | 41 | 730 | 73 | 2460 | 246 |
| beer hall | 600 | 60 | 110 | 11 | 344 | 35 | 1054 | 106 |
| butchers shop | 261 | 27 | 60 | 6 | 50 | 5 | 371 | 38 |
| cafeteria | 1443 | 145 | 200 | 20 | 370 | 37 | 2013 | 202 |
| candy store | 360 | 36 | 80 | 8 | 90 | 9 | 530 | 53 |
| coffee shop | 2060 | 206 | 260 | 26 | 590 | 59 | 2910 | 291 |
| delicatessen | 680 | 68 | 80 | 8 | 50 | 5 | 810 | 81 |
| dining room | 3020 | 302 | 420 | 42 | 930 | 93 | 4370 | 437 |
| fastfood restaurant | 920 | 92 | 150 | 15 | 330 | 33 | 1400 | 140 |
| food court | 200 | 20 | 90 | 9 | 40 | 4 | 330 | 33 |
| ice cream parlor | 160 | 16 | 50 | 5 | 60 | 6 | 270 | 27 |
| kitchen | 3300 | 330 | 400 | 40 | 990 | 99 | 4690 | 469 |
| market indoor | 800 | 80 | 150 | 15 | 210 | 21 | 1160 | 116 |
| market outdoor | 1313 | 132 | 60 | 6 | 250 | 25 | 1623 | 163 |
| picnic area | 667 | 67 | 140 | 14 | 260 | 26 | 1067 | 107 |
| pizzeria | 1120 | 112 | 370 | 37 | 600 | 60 | 2090 | 209 |
| pub indoor | 372 | 38 | 60 | 6 | 150 | 15 | 582 | 59 |
| restaurant | 4551 | 456 | 550 | 55 | 1120 | 112 | 6222 | 623 |
| supermarket | 3812 | 382 | 862 | 87 | 1423 | 143 | 6097 | 612 |
| sushi bar | 1270 | 127 | 340 | 34 | 426 | 43 | 2036 | 204 |
| Total | 29005 | 2909 | 5100 | 511 | 9287 | 932 | **43392** | **4352** |

Moreover, the "EgoFoodPlaces" dataset has some unbalanced classes. However, it is not possible to make it a balanced dataset by reducing images from other classes, since some classes have a very small number of images. The classes with few images are usually related to some food places that the users do not spend much time at them (e.g. butchers shop). In turn, some classes of a big number of images are related to places with rich visual information that refer to daily contexts (e.g. kitchen, supermarket), or places, where people spend more time (e.g. restaurant). We labeled our dataset by taking the reference classes names related to food scenes of the public Places2 dataset [32]. Initially, we chose 22 common food-related places that people often visited for our dataset. The food-related places that user visited very rarely (e.g. beer garden), were excluded from our dataset.

Finally, the 16 users recorded their period of stay (the exact time) in any food place visited during capturing the photo-streams. Afterwards, we created the events of each class by selecting the maximum correlated frames from that period. The period of stay is divided to a set of events. Each event is around 10 seconds. We select 10 seconds, because we need to keep the similarity between the consequent frames. Since our wearable camera is adjusted to capture one frame per second, one event will contain 10 consequent frames. For instance, assume a user visited a bar for 10 minutes. Thus, for a minute, we will have 6 events (60 seconds/10 seconds) and 60 events for the whole 10 minutes. The 22 classes of food places in "EgoFoodPlaces" are illustrated in Table 1.

For the training, the dataset was split into three subsets: train (70%), validation (10%) and test (20%). The images of each set were not randomly chosen to avoid taking similar images from the same events. Thus, we split the dataset based on event information in order to make the dataset more robust to train and validate the models.

### B. EXPERIMENTAL SETUP

The proposed model was implemented in Pytorch [48]: an open source deep learning library. The Adam [49] algorithm is used for model optimization. The "step" learning rate policy [50] is used with the base learning rate of 0.001 with 20 as a step value. For the LSTM cells, we used hidden size of 2048 that is similar to the output size of the MACNet feature. The number of layers is 6 and the dropout rate is 0.3. In turn, for self-attention, 22 layers are used for getting the attention score of 22 classes (number of classes in "EgoFoodPlaces"). In addition, data augmentation is applied for increasing the dataset size and variation. We performed random crop, image brightness and contrast change with 0.2 and 0.1, respectively. We also use image translation of 0.5, a random scale between 0.5 and 1.0, and random rotation of 10 degrees. The batch size is set to 64 for training with 100 epochs. The experiments are executed on NVIDIA GTX1080-Ti with 11 GB memory taking around one day to train the network. All the above parameters are used for testing the model as well.

### C. EVALUATION

In order to evaluate the proposed MACNet+SA model quantitatively, we compared it with the state-of-the-art in terms of the average $F_1$ score, and the classification accuracy rate.

The $F_1$ score can be defined as:

$$F_1 \text{ score} = 2 \times \frac{Precision.Recall}{Precision + Recall}, \tag{12}$$

**TABLE 2.** Average $F_1$ score of VGG16 [27], ResNet50 [29], InceptionV3 [47], MACNet [12] and the proposed MACNet+SA model using both validation and test sets from EgoFoodPlaces dataset.

| Categories | VGG16 | | ResNet50 | | InceptionV3 | | MACNet | | MACNet+SA | |
|---|---|---|---|---|---|---|---|---|---|---|
| dataset | val | test | val | test | val | tests | val | test | val | test |
| bakery shop | 0.77 | 0.59 | 0.72 | 0.65 | 0.77 | 0.75 | 0.74 | 0.68 | **0.85** | **0.84** |
| banquet hal | 0.71 | 0.48 | 0.62 | 0.38 | 0.73 | 0.50 | 0.64 | 0.51 | **0.75** | **0.70** |
| bar | 0.66 | 0.52 | 0.37 | 0.36 | 0.74 | 0.56 | 0.65 | 0.58 | **0.85** | **0.73** |
| beer hall | 0.77 | 0.48 | 0.92 | 0.45 | 0.91 | 0.40 | 0.95 | **0.51** | **0.96** | 0.44 |
| butchers shop | 0.71 | 0.83 | 0.72 | 0.91 | 0.72 | 0.89 | **0.79** | **0.92** | 0.73 | 0.88 |
| cafeteria | 0.61 | 0.47 | 0.64 | 0.60 | 0.70 | 0.59 | 0.78 | 0.63 | **0.94** | **0.78** |
| candy store | 0.65 | 0.59 | **0.66** | 0.63 | 0.65 | 0.57 | 0.63 | **0.64** | 0.64 | 0.58 |
| coffee shop | 0.45 | 0.71 | 0.57 | 0.71 | 0.66 | 0.68 | 0.89 | 0.75 | **0.93** | **0.87** |
| delicatessen | 0.52 | 0.62 | 0.55 | 0.73 | 0.50 | 0.64 | **0.69** | 0.56 | 0.59 | **0.75** |
| dining room | 0.62 | 0.67 | 0.71 | 0.74 | 0.73 | 0.75 | **0.92** | **0.87** | 0.87 | 0.86 |
| fastfood restaurant | 0.33 | 0.44 | 0.33 | 0.49 | 0.32 | 0.50 | **0.77** | 0.56 | 0.68 | **0.63** |
| food court | 0.64 | 0.66 | 0.63 | 0.69 | 0.70 | 0.63 | 0.82 | 0.63 | **0.86** | **0.73** |
| ice cream parlor | 0.65 | 0.64 | 0.64 | 0.60 | **0.72** | **0.69** | 0.66 | 0.64 | 0.67 | 0.65 |
| kitchen | 0.79 | 0.85 | 0.91 | 0.89 | 0.88 | 0.87 | 0.90 | 0.89 | **0.93** | **0.92** |
| market indoor | 0.53 | 0.44 | 0.56 | 0.60 | 0.40 | 0.48 | **0.81** | 0.64 | 0.76 | **0.82** |
| market outdoor | 0.42 | 0.53 | 0.37 | 0.77 | 0.39 | 0.70 | **0.61** | 0.69 | 0.48 | **0.78** |
| picnic area | 0.51 | 0.44 | 0.59 | 0.47 | 0.49 | 0.45 | 0.68 | 0.46 | **0.80** | **0.67** |
| pizzeria | 0.77 | 0.62 | 0.39 | 0.48 | 0.81 | 0.67 | 0.68 | 0.67 | **0.99** | **0.95** |
| pub indoor | 0.86 | 0.49 | **0.96** | 0.88 | 0.93 | 0.70 | 0.95 | **0.92** | 0.94 | 0.83 |
| restaurant | 0.51 | 0.47 | 0.62 | 0.46 | 0.60 | 0.51 | 0.72 | 0.55 | **0.85** | **0.66** |
| supermarket | 0.80 | 0.81 | 0.81 | 0.86 | 0.83 | 0.84 | 0.71 | 0.88 | **0.91** | **0.89** |
| sushi bar | 0.78 | 0.44 | 0.88 | 0.44 | 0.76 | 0.43 | 0.95 | 0.73 | **0**.99 | **0.88** |
| Avg. $F_1$ score | 0.66 | 0.62 | 0.68 | 0.65 | 0.72 | 0.66 | 0.79 | 0.72 | **0.86** | **0.80** |

where precision is the number of true positives divided by the total numbers of actual results, and computed as:

$$Precision = \frac{True\ positive}{True\ positive + False\ positive}, \tag{13}$$

In turn, recall is the number of true positives divided by the total number of predicted results by the classifier, and computed as:

$$Recall = \frac{True\ positive}{True\ positive + False\ negative}. \tag{14}$$

### D. RESULTS AND DISCUSSIONS

In this section, we have compared the proposed MACNet+SA model with four baseline methods: three common classification methods, VGG16 [27], ResNet50 [29], InceptionV3 [47], and the fourth one is our previous work (MACNet [12]) for both validation and test sets.

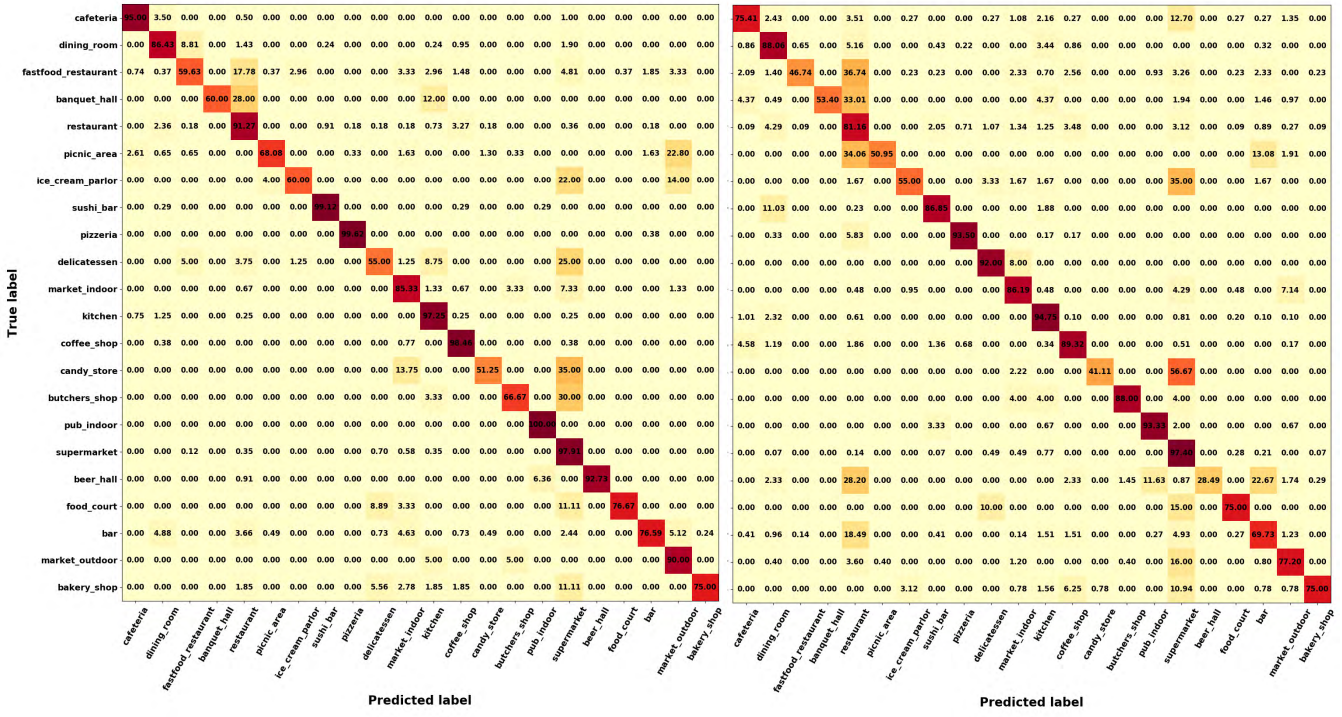
Table 2 shows the average $F_1$ score of the proposed model, MACNet+SA, and the four tested methods with the 22 classes of "EgoFoodPlaces". As shown, MACNet+SA yielded the highest average $F_1$ score of 0.86 and 0.80 for both validation and test sets, respectively. In addition, MACNet+SA achieved the highest $F_1$ score with the majority of classes in the two sets. In turn, our previous method, MACNet provided acceptable average $F_1$ score of 0.79 and 0.73 with the validation and test sets, respectively, which is higher than the other three methods, VGG16, ResNet50 and InceptionV3. The InceptionV3 achieved average $F_1$ score comparable with MACNet with 0.72, and 0.66 on the two sets. In turn, ResNet50 and VGG16 yielded similar average $F_1$ score of about 0.65.

For the validation set, with 13 out of 22 classes, MACNet+SA yielded the highest $F_1$ score. In turn, with 6 out of 9 remaining classes, the predecessor MACNet achieved the highest $F_1$ score. While for candy store and pub indoor classes ResNet50 had the highest $F_1$ score. For the ice cream parlor class, InceptionV3 model yielded the highest $F_1$ score. In turn, VGG16 achieved the lowest $F_1$ score among the five tested methods for all classes.

For the test set, the proposed MACNet+SA yielded the highest $F_1$ score with 16 out of 22 classes. In turn, the predecessor model MACNet achieved the highest $F_1$ score in 5 out of 6 remaining classes. In turn, the InceptionV3 yielded the highest $F_1$ score for the ice cream parlor class. In addition, both VGG16 and ResNet50 models achieved lower $F_1$ score than the rest of the tested models for all classes.

The proposed MACNet+SA yielded an average improvement of 7% and 8% in terms of the average $F_1$ score with the validation and test sets, respectively in a comparison of the second best state-of-the-art *i.e.*, its predecessor MACNet. In some places like bar, cafeteria, picnic area, pizzeria and other places that need a sequence of images to describe them, MACNet+SA yielded a significant improvement of more than 10%. However, with some classes, such as butchers shop, dining room, market indoor and market outdoor, MACNet provided higher results than MACNet+SA showing that these type of places might not need to describe them with a sequence of images, and still images are able to describe these places.

In turn, Table 3 shows a comparison between the proposed MACNet+SA model with MACNet VGG16, ResNet50 and InceptionV3 in terms of Top-1 and Top-5 classification




(a) (b)

**FIGURE 7.** The confusion matrices of (a) validation and (b) test sets of the EgoFoodPlaces dataset for evaluating our propose model.

**TABLE 3.** Average Top-1 and Top-5 classification accuracy of VGG16 [27], ResNet50 [29], InceptionV3 [47], MACNet [12] and the proposed MACNet+SA model using both validation and test sets from EgoFoodPlaces dataset.

| Models | Validation | | Test | |
|---|---|---|---|---|
| | Top-1 | Top-5 | Top-1 | Top-5 |
| VGG16 | 0.66 | 0.87 | 0.62 | 0.86 |
| ResNet50 | 0.68 | 0.91 | 0.65 | 0.90 |
| InceptionV3 | 0.72 | 0.91 | 0.66 | 0.88 |
| MACNet | 0.79 | 0.90 | 0.72 | 0.89 |
| **MACNet+SA** | **0.86** | **0.93** | **0.80** | **0.92** |

accuracy rates on both validation and test sets. It shows that MACNet+SA achieved the highest Top-1 and Top-5 accuracy rates with the two sets. Regarding the validation set, MACNet+SA yielded an improvement of 7% and 2% in terms of top-1 and top-5 rates, respectively, higher than the MACNet model achieving the highest classification rate among the four test models. In turn, for the test set, MACNet+SA yielded an improvement of 8% and 2% with top-1 and top-5 rates, respectively.

Furthermore, Figure 7 shows a confusion matrix of the 22 classes of the EgoFoodPlaces dataset with the validation and test sets. The confusion matrix in Figure 7-(a) shows that the proposed model, MACNet+SA, with the validation set, was able to correctly classify the food-places events in most of the classes. However, it misclassifies events from a class to another. For example, MACNet+SA misclassifies 17.78% of fastfood restaurant events to the restaurant class, in addition, 22.80% of picnic area events are misclassified with the outdoor market class, and 22% and 14% of ice cream parlour samples are misclassified with the supermarket and outdoor market classes, respectively. The confusion matrix also shows that 25% of delicatessen events are misclassified with the supermarket class, 35% of candy store samples are misclassified with the supermarket class, 28% of banquet hall samples are misclassified with the restaurant class, and 30% of butcher shop events are misclassified with the supermarket class. The confusion matrix in Figure 7 (b) shows that the proposed classification model with the test set misclassifies events from classes to restaurant, supermarket and bar classes. It shows 36.74%, 33.01%, 33.01%, 34.06%, and 18.49% of the events of the fastfood restaurant, banquet hall, picnic area, beer hall and bar classes are misclassified to the restaurant class. In addition, the confusion matrix shows 12.70%, 35%, 56.57%, 15%, 16%, and 10.94% of cafeteria, icecream parlour, candy store, food court, market outdoor and bakery shop events are misclassified with the supermarket class. Similarly, 13.08%, and 22.67% of picnic area and beer hall events, respectively, are misclassified with the bar class. However, for all of these misclassifications events, there is a lot of similarity between their scenes in terms of the context and objects. Even, humans prone to weakly recognize such places many times.

Figure 8 shows examples of correct and incorrect predictions by the proposed MACNet+SA model with the "EgoFoodPlaces" dataset. The first, third, fifth and seventh rows show that the proposed MACNet+SA model is able to properly predict all images of events of the dining room,

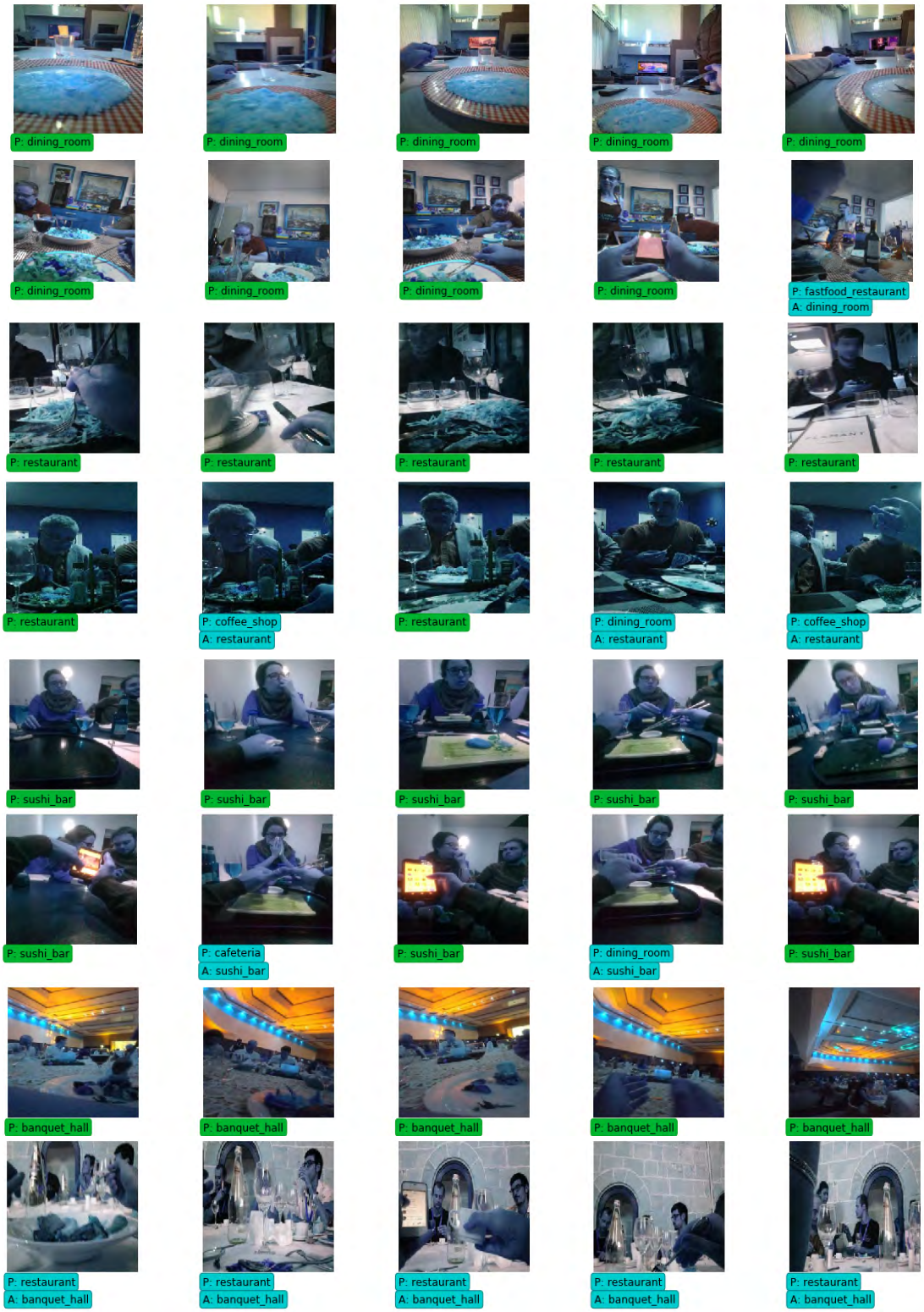



**FIGURE 8. Examples of correct and incorrect predictions of MACNet+SA model with the input event (a sequence of images) of the validation set.**

restaurant, sushi bar and banquet hall classes, respectively. In turn, second, fourth and sixth and last rows show incorrect predictions examples, in which one image or more of the dining room, restaurant, sushi bar and banquet hall events are misclassified. In the second row, images in first, second, third and fourth columns are correctly classified as a dining room class; whereas, the images in the last image is misclassified as fastfood restaurant. In the fourth row, the restaurant class is


**FIGURE 9.** Examples of the resulting predictions (from Top-1 to Top-5) of the proposed MACNet+SA model using validation dataset, where GT is the ground-truth label of the predicted class.

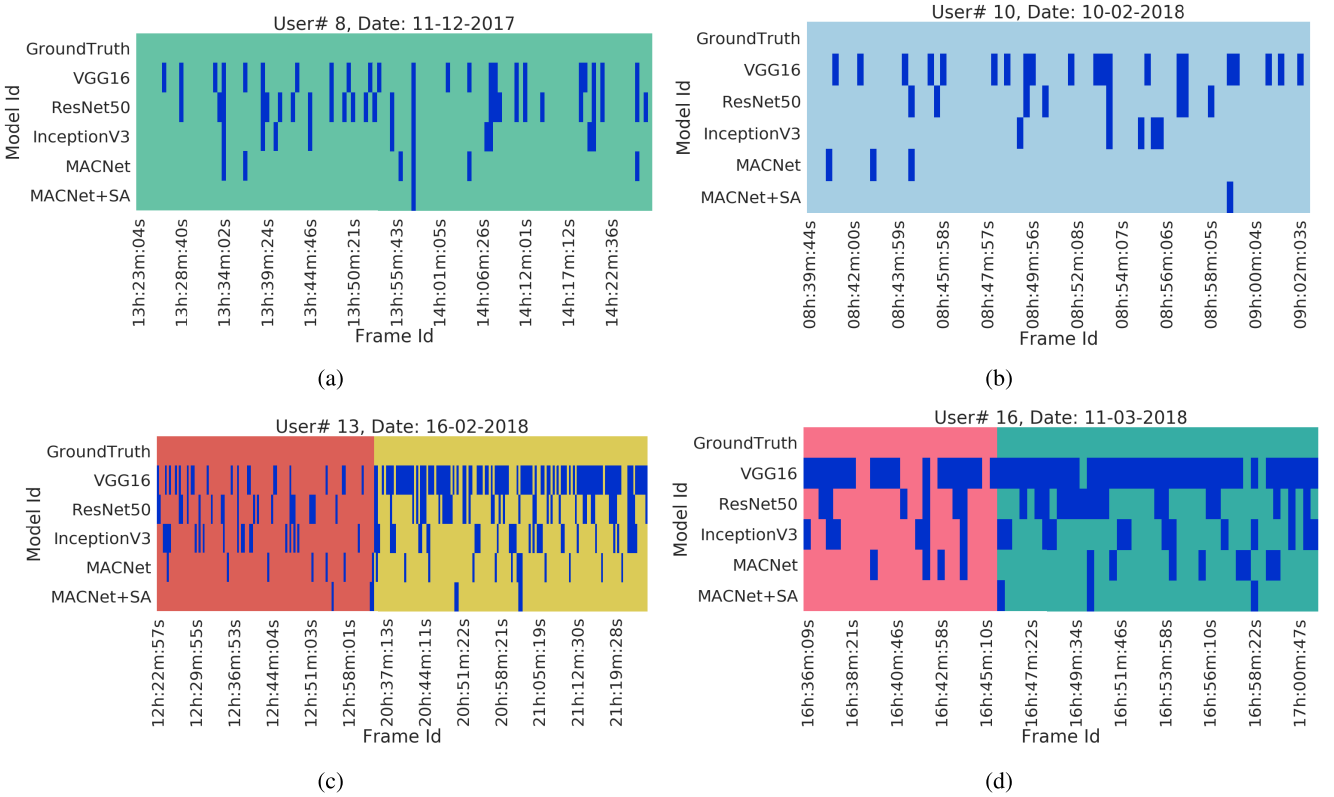



**FIGURE 10.** Resulted food places classification with four periods of stay in six food places (coffee shop, bakery shop, food court, sushi bar, kitchen and dining room) captured by four different users (users 8, 10, 13 and 16 of the EgoFoodPlaces dataset) in four different days from the validation set.

correctly predicted with first and third images, while second and last images are misclassified as a coffee shop and the dining room, respectively. However, the Top-2 prediction is the correct class, restaurant. In the sixth row, the sushi bar class is correctly predicted with the first, third and last images. In turn, the second and fourth images are misclassified as cafeteria and dining room classes, respectively. In the last row, all images of a banquet hall event are predicted as a restaurant class. However, with all images, the Top-2 prediction is the banquet hall class.

Figure 9 shows examples of predicted Top-1 to Top-5 accuracy. The first row shows that cafeteria, kitchen and restaurant images are properly classified with Top-1 classification accuracy rates of 93.06%, 84.99% and 89.67%, respectively. In turn, the second row shows the proposed MACNet+SA model wrongly predicted restaurant, dining room and

fastfood restaurant classes with the Top-1 accuracy. However, these classes barely appeared in Top-5 accuracy with a restaurant in Top-2, dining room in Top-3 and fastfood restaurant in Top-5, with a classification accuracy of 39.60%, 3.65% and 11.36%, respectively.

Figure 10 shows four period of stays in six food places captured by four different users (users 8, 10, 13 and 16 of the ''EgoFoodPlaces'' dataset) in four different days. The user 8 visited a coffee shop for 59 minutes, and user 10 visited bakery shop for 22 minutes. In addition, the third and fourth users visited two different food places: food court and sushi bar for user 13, whereas kitchen and dining room for user 16. All events during each period were tested with the proposed MACNet+SA model. For instance, for user 8, he spent 59 min in a coffee shop, we divided it into 354 events. In turn, for user 16, 54 events were included in his first stay in kitchen (i.e. 9 minutes), and 72 events during his stay inside a dining room (i.e. 12 minutes). One can notice that the proposed MACNet+SA model yielded the lowest misclassification rates in the four sequences of events. With the events sequences of user 8 and 10 in coffee and bakery shops, respectively, the proposed MACNet+SA model misclassified only one event per every sequence. In the third events sequence of user 13, MACNet+SA misclassified two events in the food court and five events in the sushi bar. In turn, for user 16, the proposed model properly predicted all events of the kitchen. However, it misclassified three events in the dining room. Supporting the aforementioned results, the MACNet model provides the second rank after the MACNet+SA with misclassification of 6, 3, 19, and 12 events with user 8, 10, 13 and 16, respectively. In turn, the VGG16 provided the worst classification rate among the all tested models.

When considering capturing images of daily life of persons and their environment, wearable devices with first-person cameras can raise some privacy concerns, since they can capture extremely private moments and sensitive information of the user. There are five steps of data privacy consideration in life-logging [51]: capture, storage, processing, access and publication. The first three phases have no human involvement. In the final two stages, the data can be accessed by humans. To deal with the private issues in real-life applications, the images can be online processed with the trained model with only storing the logging information without any confidential data and avoiding to store the images during the logging process. Also, the user can handle the system with mobile apps to turn off in private moments and turn on when entering to food places. Taking this viewpoint, we consider that the right to privacy in terms of life-logging refers to *the right to choose the composition and the usage of your life-log and the right to choose what happens to your representation in the life-logs of others* [51]

## V. CONCLUSIONS

In this paper, we proposed a deep food places classification system, MACNet+SA, for egocentric photo-streams captured during a day. The main purpose of this classification system is to later generate a dietary report to analyze people's food intake and help them control their unhealthy dietary habits. The proposed deep model is based on a self-attention model with the MACNet model proposed in [12]. The MACNet model used atrous convolutional networks to classify still images. However, the proposed model classifies a sequence of images (called events) to get relevant temporal information about the food places. Image-level features are extracted by the MACNet model. The LSTM cells with a self-attention mechanism merge the temporal information of the sequence of the input images. The quantitative and qualitative results show that the proposed MACNet+SA model is able to outperform state of the art classification methods, as VGG16, ResNet50, InceptionV3 and MACNet. MACNet+SA on the dataset, EgoFoodPlaces, yields an average $F_1$ score of 86% and 80% on validation and test set, respectively. In addition, it yields a Top-1 accuracy of 86% and 80%, and a Top-5 accuracy of 93% and 92% on the validation and test sets, respectively. Future work aims at developing a mobile application based on the MACNet+SA model that integrates an egocentric camera with a personal mobile device to create a dietary report to keep a track on our eating behavior or routine for following a healthy diet.

## REFERENCES

[1] C. M. Hales, C. D. Fryar, M. D. Carroll, D. S. Freedman, and C. L. Ogden, ''Trends in obesity and severe obesity prevalence in US youth and adults by sex and age, 2007-2008 to 2015-2016,'' *Jama*, vol. 319, no. 16, pp. 1723–1725, 2018.
[2] M. Peralta, M. Ramos, A. Lipert, J. Martins, and A. Marques, ''Prevalence and trends of overweight and obesity in older adults from 10 European countries from 2005 to 2013,'' *Scand. J. Public Health*, vol. 46, no. 5, pp. 522–529, 2018.
[3] A. B. Keys, ''Overweight, obesity, coronary heart disease, and mortality,'' *Nutrition Rev.*, vol. 38, no. 9, pp. 297–307, 1980.
[4] E. A. Finkelstein, J. G. Trogdon, J. W. Cohen, and W. Dietz, ''Annual medical spending attributable to obesity: Payer-and service-specific estimates,'' *Health Affairs*, vol. 28, no. 5, pp. w822–w831, 2009.
[5] S. Cuschieri and J. Mamo, ''Getting to grips with the obesity epidemic in Europe,'' *SAGE Open Med.*, vol. 4, Sep. 2016, Art. no. 2050312116670406.
[6] M. Bolaños, M. Dimiccoli, and P. Radeva, ''Toward storytelling from visual lifelogging: An overview,'' *IEEE Trans. Hum.-Mach. Syst.*, vol. 47, no. 1, pp. 77–90, Feb. 2017.
[7] M. Aghaei, M. Dimiccoli, C. C. Ferrer, and P. Radeva, ''Towards social pattern characterization in egocentric photo-streams,'' *Comput. Vis. Image Understanding*, vol. 171, pp. 104–117, Jun. 2018.
[8] M. Aghaei, M. Dimiccoli, and P. Radeva, ''Towards social interaction detection in egocentric photo-streams,'' *Proc. SPIE*, vol. 9875, Dec. 2015, Art. no. 987514.
[9] E. R. Grimm and N. I. Steinle, ''Genetics of eating behavior: Established and emerging concepts,'' *Nutrition Rev.*, vol. 69, no. 1, pp. 52–60, 2011.
[10] E. Kemps, M. Tiggemann, and S. Hollitt, ''Exposure to television food advertising primes food-related cognitions and triggers motivation to eat,'' *Psychol. & health*, vol. 29, no. 10, pp. 1192–1205, 2014.
[11] R. A. de Wijk, I. A. Polet, W. Boek, S. Coenraad, and J. H. Bult, ''Food aroma affects bite size,'' *Flavour*, vol. 1, no. 1, p. 3, 2012.
[12] M. M. K. Sarker, H. A. Rashwan, E. Talavera, S. F. Banu, P. Radeva, and D. Puig, ''MACNet: Multi-scale atrous convolution networks for food places classification in egocentric photo-streams,'' in *Proc. Eur. Conf. Comput. Vis.* Munich, Germany: Springer, Sep. 2018, pp. 423–433.
[13] A. Oliva and A. Torralba, ''Scene-centered description from spatial envelope properties,'' in *Proc. Int. Workshop Biol. Motivated Comput. Vis.* Tübingen, Germany: Springer, Nov. 2002, pp. 263–272.
[14] J. Luo and M. Boutell, ''Natural scene classification using overcomplete ICA,'' *Pattern Recognit.*, vol. 38, no. 10, pp. 1507–1519, 2005.

[15] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[16] J. Yu, D. Tao, Y. Rui, and J. Cheng, "Pairwise constraints based multiview features fusion for scene classification," *Pattern Recognit.*, vol. 46, no. 2, pp. 483–496, 2013.
[17] L.-J. Li, R. Socher, and L. Fei-Fei, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2036–2043.
[18] J. Qin and N. H. C. Yung, "Scene categorization via contextual visual words," *Pattern Recognit.*, vol. 43, no. 5, pp. 1874–1888, 2010.
[19] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Learning hierarchical models of scenes, objects, and parts," in *proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2, Oct. 2005, pp. 1331–1338.
[20] N. M. Elfiky, F. S. Khan, J. van de Weijer, and J. González, "Discriminative compact pyramids for object and scene recognition," *Pattern Recognit.*, vol. 45, no. 4, pp. 1627–1636, 2012.
[21] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.
[22] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb, "Reconfigurable models for scene recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2775–2782.
[23] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 2169–2178.
[24] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 413–420.
[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[26] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[27] K. Simonyan and A. Zisserman. (Sep. 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[28] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[30] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.
[31] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3485–3492.
[32] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
[33] L. Zheng, S. Wang, F. He, and Q. Tian. (2014). "Seeing the big picture: Deep embedding with contextual evidences." [Online]. Available: https://arxiv.org/abs/1406.0132
[34] R. Wu, B. Wang, W. Wang, and Y. Yu, "Harvesting discriminative meta objects with deep CNN features for scene classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1287–1295.
[35] A. Furnari, G. M. Farinella, and S. Battiato, "Temporal segmentation of egocentric videos to highlight personal locations of interest," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 474–489.
[36] A. Furnari, G. M. Farinella, and S. Battiato, "Recognizing personal locations from egocentric videos," *IEEE Trans. Hum.-Mach. Syst.*, vol. 47, no. 1, pp. 6–18, Feb. 2017.
[37] M. M. K. Sarker *et al.*, "Foodplaces: Learning deep features for food related scene understanding," in *Proc. Recent Adv. Artif. Intell. Res. Develop., 20th Int. Conf. Catalan Assoc. Artif. Intell. (CCIA)*, 2017, pp. 156–165.
[38] M. Sarker, M. Jabreel, and H. A. Rashwan, "Cuisinenet: food attributes classification using multi-scale convolution network," in *Proc. Artif. Intell. Res. Develop., Current Challenges, New Trends Appl. (CCIA)*, vol. 308, 2018, p. 365.
[39] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.
[40] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[41] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[42] C. Hori *et al.*, "Attention-based multimodal fusion for video description," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4203–4212.
[43] M. Jabreel, F. Hassan, S. Abdulwahab, and A. Moreno, "Recurrent neural conditional random fields for target identification of tweets," in *Proc. CCIA*, Oct. 2017, pp. 66–75.
[44] M. Jabreel, F. Hassan, and A. Moreno, "Target-dependent sentiment analysis of tweets using bidirectional gated recurrent neural networks," in *Proc. Adv. Hybridization Intell. Methods*. Springer, 2018, pp. 39–55.
[45] Z. Lin *et al.*. (Mar. 2017). "A structured self-attentive sentence embedding." [Online]. Available: https://arxiv.org/abs/1703.03130
[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[47] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.
[48] A. Paszke, S. Gross, S. Chintala, and G. Chanan, "Pytorch," Tech. Rep., 2017.
[49] D. P. Kingma and J. Ba. (Dec. 2014). "Adam: A method for stochastic optimization," [Online]. Available: https://arxiv.org/abs/1412.6980
[50] A. Schoenauer-Sebag, M. Schoenauer, and M. Sebag. (2017). "Stochastic gradient descent: Going as fast as possible but not faster." [Online]. Available: https://arxiv.org/abs/1709.01427
[51] C. Gurrin, R. Albatal, H. Joho, and K. Ishii, *A Privacy by Design Approach to Lifelogging*. Amsterdam, The Netherlands: IOS Press, 2014, pp. 49–73.

**MD. MOSTAFA KAMAL SARKER** received the B.S. degree from the Shahjalal University of Science and Technology, Sylhet, Bangladesh, in 2009, and the M.S. degree from Chonbuk National University, Jeonju, South Korea, in 2013, supported by the Korean government "Brain Korea21 (BK21)" Scholarship Program. He is currently pursuing the Ph.D. degree with the Intelligent Robotics and Computer Vision Group, Department of Computer Engineering and Mathematics Security, Rovira i Virgili Univerisity, where he has been a Predoctoral Researcher, since 2016. From 2013 to 2016, he was a Researcher on a project from the National Research Foundation of Korea (NRF) that is funded by the Ministry of Education of South Korea. His research interests include the areas of image processing, pattern recognition, computer vision, machine learning, deep learning, egocentric vision, and visual lifelogging.

**HATEM A. RASHWAN** received the B.S. degree in electrical engineering and the M.Sc. degree in computer science from South Valley University, Aswan, Egypt, in 2002 and 2007, respectively, and the Ph.D. degree in computer vision from Rovira i Virgili University, Tarragona, Spain, in 2014. From 2004 to 2009, he was with the Electrical Engineering Department, Aswan Faculty of Engineering, South Valley University, Egypt, as a Lecturer. In 2010, he joined the Intelligent Robotics and Computer Vision Group, Department of Computer Science and Mathematics, Rovira i Virgili University, where he was a Research Assistant with the IRCV Group, in 2014. From 2014 to 2017, he was a Researcher in vortex with IRIT-CNRS, INP-ENSEEIHT, University of Toulouse, Toulouse, France. Since 2017, he has been a Beatriu de Pinos Researcher with DEIM, Universitat Rovira i Vigili. His research interests include image processing, computer vision, pattern recognition, and machine learning.

**FARHAN AKRAM** received the B.Sc. degree in computer engineering from the COMSATS Institute of Information Technology, Islamabad, Pakistan, in 2010, the M.Sc. degree in computer science with a major in application software from Chung-Ang University, Seoul, South Korea, in 2013, and the Ph.D. degree in computer engineering and mathematics from Rovira i Virgili University, Tarragona, Spain, in 2017. He joined the Imaging Informatics Division, Bioinformatics Institue, A*STAR, Singapore, in 2017, as a Postdoctoral Research Fellow and still working there. His current research interests include medical image analysis, image processing, computer vision, and deep learning.

**ESTEFANIA TALAVERA** received the B.Sc. degree in electronic engineering from Balearic Islands University, in 2012, and the M.Sc. degree in biomedical engineering from the Polytechnic University of Catalonia, in 2014. She is currently pursuing the Ph.D. degree with the University of Barcelona and the University of Groningen. Her research interests include lifelogging and health applications.

**SYEDA FURRUKA BANU** received the B.Sc. degree in statistics from the Shahjalal University of Science and Technology, Sylhet, Bangladesh, in 2011. She is currently pursuing the M.S. degree in technology and engineering management with Rovira i Virgili University, Spain. Her research interests include statistical analysis, machine learning, and social and organizational analysis.

**PETIA RADEVA** is currently a Senior Researcher and an Associate Professor with the University of Barcelona. She is also the Head of Computer Vision with the University of Barcelona Group and the Medical Imaging Laboratory, Computer Vision Center. Her present research interests include the development of learning-based approaches for computer vision, egocentric vision, and medical imaging.

**DOMENEC PUIG** received the M.S. and Ph.D. degrees in computer science from the Polytechnic University of Catalonia, Barcelona, Spain, in 1992 and 2004, respectively. In 1992, he joined the Department of Computer Science and Mathematics, Rovira i Virgili University, Tarragona, Spain, where he is currently an Associate Professor. Since 2006, he has been the Head of the Intelligent Robotics and Computer Vision Group, Rovira i Virgili University. His research interests include image processing, texture analysis, perceptual models for image analysis, scene analysis, and mobile robotics.